Online Appendix for "The Consumer Welfare Effects of Online Ads: Evidence from a 9-Year Experiment"

Erik Brynjolfsson, Avinash Collis, Daniel Deisenroth, Haritz Garro, Daley Kutzman, Asad Liaqat, and Nils Wernerfelt

This Appendix contains four sections: first, we provide evidence on covariate balance across our treatment and control sample; second, we show the output behind our results in the main text and perform robustness checks; third, we provide additional details on the survey flow and execution; and finally, we discuss more detail on how we weighted our results.

A.1 Covariate Balance

Below we provide evidence on whether the experiment randomization and sample recruitment led to users in our treatment and control samples being comparable. We present two balance tables: Table A.1 contains the data from our raw, unweighted survey sample from users who were either in the ads or no ads condition; Table A.2 is the reweighted version of that.

We include 18 demographics and pre-determined characteristics in our data, showing the average difference between the two groups, significance level of the difference, the difference as a percent of the ads baseline (in absolute value), and the share missing. We note there are significant differences between the two groups, with individual demographics and with a joint significance test for the unweighted and weighted tables ($F = 7.922$, $p < 0.001$ and $F = 9.388$, $p < 0.001$, respectively).

As mentioned in the main text, however, we highlight that the magnitude of the differences across the demographics is small for both the unweighted and weighted sample. The average difference as a percent of the ads baseline is 2.5% in the weighted table and 2.4% in the unweighted one. Further, after reweighting, only female and Facebook tenure are significantly different, and rerunning all our analyses controlling for both of these variables yields no material difference in the results.

Table A.1: Covariate balance table (53,083 survey respondents).

	Difference (Ads - No Ads)	***p*-value**	**Difference as fraction of Ads Baseline (abs. value)**	**% Missing**
Age (years)	0.3759	0.003***	0.0104	0%
Age 18-24	-0.0031	0.419	0.0178	0%
Age 25-34	-0.0094	0.050**	0.0263	0%
Age 35-44	0.0013	0.754	0.0055	0%
Age 45-54	0.0073	0.030**	0.0537	0%
Age 55-64	-0.0001	0.957	0.0021	0%
Age 65+	0.0039	0.015**	0.1333	0%
Female	-0.0283	<0.001***	0.0742	0%
Homeowner	0.0006	0.920	0.0012	28%
Finished HS	0.0081	0.005***	0.0087	23%
Finished college	0.0231	<0.001***	0.0475	23%
Has profile photo	-0.0155	<0.001***	0.0162	0%
Primary Phone OS is iOS	0.0008	0.871	0.002	0%
Primary Phone OS is Android	-0.0012	0.805	0.002	0%
Facebook tenure (days)	73.0325	<0.001***	0.0197	0%
Contact email confirmed	0.0088	1.000	0.0095	0%
Day of week born	-0.0144	0.465	0.0036	3%
Month born	-0.0297	0.396	0.0046	3%

Note: This table shows the unweighted comparison between the ads and no ads groups. All variables are binary except age (years), Facebook tenure (days), day of week born (1-7), and month born (1-12). The variables Homeowner, Finished HS (High School), and Finished college were self-reported in our survey, while the other variables were obtained from platform data. The first numeric column shows the raw difference between the averages in the two groups; the second numeric column shows the p-value on the difference; the third numeric column shows the raw difference as a fraction of the ads group baseline; and the last column shows the share of our data that were missing entries for that demographic. A joint significance test finds evidence of a significant difference between the groups ($F = 7.922$, $p < 0.001$), though as mentioned in the main text, the magnitudes of the differences remain small. ** denotes $p < 0.05$, *** denotes $p < 0.01$.

Table A.2: Reweighted covariate balance table (53,083 survey respondents).

	Difference (Ads - No Ads)	***p*-value**	**Difference as fraction of Ads Baseline (abs. value)**	**% Missing**
Age (Years)	-0.157	0.473	0.0042	0%
Age 18 -24	0.008	0.141	0.0497	0%
Age 25-34	-0.007	0.321	0.0211	0%
Age 35-44	-0.002	0.735	0.0091	0%
Age 45-54	-0.002	0.694	0.0118	0%
Age 55-64	0.007	0.127	0.0897	0%
Age 65+	-0.004	0.266	0.0976	0%
Female	0.039	<0.001***	0.0816	0%
Homeowner	-0.001	0.897	0.0020	28%
Finished HS	0.000	0.922	0.0000	23%
Finished College	-0.007	0.401	0.0158	23%
Has a profile picture	0.000	0.906	0.000	0%
Primary Phone OS is iOS	-0.006	0.447	0.0132	0%
Primary Phone OS is Android	0.004	0.567	0.0074	0%
Facebook tenure (days)	-107.736	<0.001***	0.0307	0%
Contact email confirmed	0.003	0.531	0.0033	0%
Day of week born	0.024	0.414	0.0060	3%
Month born	-0.025	0.628	0.0038	3%

Note: This table shows the comparison between the ads and no ads groups, now weighted as discussed in the main text and Appendix A.4. A joint significance test finds evidence of a significant difference between the groups ($F = 9.388$, $p < 0.001$), though as mentioned in the main text, the magnitudes of the differences remain small. *** denotes p < 0.01.

A.2 Main Output and Robustness Checks

In this section, we report the output behind the Figures in the main text as well as from different robustness checks.

In Tables A.3 and A.4 we provide the data behind our main figure, Figure 4. We note the rejection rates are only significantly different twice across all the comparisons, with neither surviving a Bonferroni correction.

Table A.3: Unweighted rejection rates across groups at each offer level (Figure 4a).

Offer Value	Ads Average Rejection Rate	No Ads Average Rejection Rate	Ads Std. Error	No Ads Std. Error	p-value
$5	0.814	0.817	0.00584	0.0100	0.7956
$10	0.735	0.735	0.00672	0.0113	1.0000
$20	0.605	0.572	0.00737	0.0131	0.0281**
$30	0.525	0.508	0.00750	0.0128	0.2518
$40	0.434	0.442	0.00756	0.0128	0.5905
$50	0.376	0.367	0.00727	0.0125	0.5337
$65	0.324	0.322	0.00708	0.0122	0.8872
$80	0.286	0.269	0.00680	0.0114	0.2003
$100	0.241	0.255	0.00635	0.0113	0.2801

Table A.4: Weighted rejection rates across groups at each offer level (Figure 4b).

Offer Value	Ads Average Rejection Rate	No Ads Average Rejection Rate	Ads Std. Error	No Ads Std. Error	p-value
$5	0.806	0.819	0.0117	0.0127	0.4515
$10	0.741	0.746	0.0131	0.0140	0.7943
$20	0.631	0.597	0.0145	0.0164	0.1204
$30	0.512	0.530	0.0150	0.0159	0.4102
$40	0.449	0.450	0.0148	0.0164	0.9639
$50	0.382	0.396	0.0141	0.0159	0.5100
$65	0.342	0.356	0.0144	0.0158	0.5125
$80	0.300	0.286	0.0136	0.0147	0.4845
$100	0.253	0.287	0.0127	0.0151	0.0849*

In Table A.5 we include the output from our main specification (Equation 1) for the total sample and each subgroup we consider. In Table A.6 we repeat the analysis but adding a control for tenure on Facebook; the results do not change meaningfully.

Table A.5: Logistic regression output for the whole sample and each subgroup. Dependent variable is a user-level rejection dummy.

	Total	US	EU	GB	MX	Bottom TS Tercile	Middle TS Tercile	Upper TS Tercile	Bottom Tenure Tercile	Middle Tenure Tercile	Upper Tenure Tercile
Offer	-0.863***	-0.870***	-0.880***	-1.091***	-0.984***	-0.914***	-0.923***	-0.824***	-0.823***	-0.884***	-0.940***
	(0.0278)	(0.0573)	(0.0483)	(0.0752)	(0.0554)	(0.0542)	(0.0488)	(0.0438)	(0.0451)	(0.0485)	(0.0510)
Ads	0.0355	-0.261	0.0231	0.134	0.286	-0.290	0.0235	0.388*	0.109	-0.252	0.189
	(0.138)	(0.307)	(0.201)	(0.338)	(0.278)	(0.252)	(0.246)	(0.224)	(0.220)	(0.243)	(0.261)
Ads X Offer	-0.0175	0.0478	-0.0123	-0.0439	-0.102	0.0806	-0.00952	-0.103*	-0.0624	0.0568	-0.0404
	(0.0382)	(0.0823)	(0.0550)	(0.0916)	(0.0804)	(0.0717)	(0.0678)	(0.0612)	(0.0619)	(0.0670)	(0.0709)
Constant	2.988***	3.454***	3.075***	3.834***	2.662***	2.780***	3.277***	3.050***	2.612***	3.084***	3.572***
	(0.1000)	(0.214)	(0.177)	(0.277)	(0.191)	(0.189)	(0.176)	(0.160)	(0.160)	(0.175)	(0.187)
Mean Rejection Rate for No Ads Group	0.499	0.587	0.504	0.496	0.339	0.427	0.513	0.538	0.448	0.503	0.565
N	53,083	5,378	24,960	5,473	6,099	17,690	17,691	17,691	17,678	17,690	17,709

Table A.6: Rerunning Table A.5, including user-level control for tenure on Facebook.

	Total	US	EU	GB	MX	Bottom TS Tercile	Middle TS Tercile	Upper TS Tercile	Bottom Tenure Tercile	Middle Tenure Tercile	Upper Tenure Tercile
Offer	-0.874***	-0.873***	-0.881***	-1.095***	-0.995***	-0.926***	-0.928***	-0.840***	-0.823***	-0.892***	-0.941***
	(0.028)	(0.058)	(0.048)	(0.075)	(0.056)	(0.055)	(0.049)	(0.045)	(0.045)	(0.049)	(0.051)
Ads	0.015	-0.262	0.019	0.138	0.298	-0.300	0.010	0.358	0.109	-0.274	0.187
	(0.139)	(0.308)	(0.201)	(0.338)	(0.278)	(0.253)	(0.247)	(0.226)	(0.220)	(0.243)	(0.261)
Ads X Offer	-0.016	0.047	-0.011	-0.046	-0.104	0.081	-0.011	-0.100	-0.063	0.062	-0.040
	(0.038)	(0.082)	(0.055)	(0.092)	(0.081)	(0.072)	(0.068)	(0.062)	(0.062)	(0.067)	(0.071)
Constant	3.036***	3.467***	3.078***	3.848***	2.692***	2.831***	3.289***	3.123***	2.673***	2.825***	3.641***
	(0.101)	(0.215)	(0.177)	(0.278)	(0.191)	(0.191)	(0.176)	(0.163)	(0.166)	(0.184)	(0.231)
Mean Rejection Rate for No Ads Group	0.499	0.586	0.504	0.496	0.339	0.426	0.513	0.537	0.448	0.503	0.565
Tenure control	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N	53,077	5,375	24,957	5,473	6,099	17,689	17,689	17,688	17,678	17,690	17,709

In Table A.7 we report the numbers behind Figure 5a, b, and c in the main text. In Table A.8 we include those numbers when tenure on Facebook is added as a control (again, we find no material differences in the results).

Table A.7: Numeric values behind Figure 5, with p-values on the differences. (Note: TS refers to 'Time Spent'; sample sizes as per the columns in Table A.5)

	Total	**US**	**EU**	**GB**	**MX**
Ads	\$31.04 (\$29.53, \$32.56)	\$48.53 (\$42.76, \$54.30)	\$32.23 (\$30.62, \$33.83)	\$32.98 (\$30.53, \$35.42)	\$15.10 (\$13.47, \$16.74)
No Ads	\$31.95 (\$30.26, \$33.63)	\$52.9 (\$47.12, \$58.69)	\$32.95 (\$29.85, \$36.05)	\$33.57 (\$30.10, \$37.04)	\$14.95 (\$13.09, \$16.81)
Δ Ads - No Ads	-\$0.90 (-\$3.16, \$1.35) p = 0.43	-\$4.38 (-\$12.70, \$3.95) p = 0.30	-\$0.73 (-\$4.26, \$2.80) p = 0.69	-\$0.59 (-\$4.89, \$3.70) p = 0.79	\$0.15 (-\$2.35, \$2.66) p = 0.90

	Bottom Tenure Tercile	**Middle Tenure Tercile**	**Upper Tenure Tercile**
Ads	\$21.64 (\$19.76, \$23.51)	\$30.71 (\$28.00, \$33.41)	\$46.29 (\$42.62, \$49.96)
No Ads	\$23.92 (\$21.72, \$26.12)	\$32.78 (\$29.91, \$35.66)	\$44.65 (\$40.50, \$48.80)
Δ Ads - No Ads	-\$2.29 (-\$5.20, \$0.63) p = 0.12	-\$2.08 (-\$6.10, \$1.95) p = 0.31	\$1.64 (-\$3.95, \$7.23) p = 0.57

	Bottom TS Tercile	**Middle TS Tercile**	**Upper TS Tercile**
Ads	\$19.86 (\$17.73, \$21.99)	\$34.40 (\$31.52, \$37.28)	\$40.75 (\$37.78, \$43.71)
No Ads	\$20.96 (\$18.72, \$23.20)	\$34.78 (\$31.73, \$37.84)	\$40.47 (\$36.94, \$43.99)
Δ Ads - No Ads	-\$1.10 (-\$4.15, \$1.95) p = 0.48	-\$0.38 (-\$4.47, \$3.70) p = 0.85	\$0.28 (-\$4.39, \$4.95) p = 0.91

Table A.8: Numeric values behind Figure 5, adding control for tenure (sample sizes as per the columns in Table A.6).

	Total	**US**	**EU**	**GB**	**MX**
Ads	\$30.73 (\$29.20, \$32.25)	\$48.40 (\$42.88, \$53.93)	\$32.21 (\$30.79, \$33.64)	\$32.97 (\$30.48, \$35.47)	\$15.16 (\$13.44, \$16.89)
No Ads	\$32.20 (\$30.54, \$33.85)	\$53.14 (\$46.92, \$59.36)	\$32.96 (\$30.11, \$35.82)	\$33.61 (\$30.02, \$37.20)	\$14.95 (\$13.14, \$16.76)
Δ Ads - No Ads	-\$1.47 (-\$3.69, \$0.76) p = 0.197	-\$4.74 (-\$12.63, \$3.16) p = 0.239	-\$0.75 (-\$3.93, \$2.44) p = 0.646	-\$0.64 (-\$5.07, \$3.80) p = 0.778	\$0.22 (-\$2.27, \$2.70) p = 0.864

	Bottom Tenure Tercile	**Middle Tenure Tercile**	**Upper Tenure Tercile**
Ads	\$21.63 (\$19.89, \$23.38)	\$30.61 (\$27.76, \$33.46)	\$46.30 (\$42.38, \$50.22)
No Ads	\$23.92 (\$21.71, \$26.14)	\$32.85 (\$29.91, \$35.80)	\$44.63 (\$40.71, \$48.55)
Δ Ads - No Ads	-\$2.29 (-\$5.05, \$0.47) p = 0.104	-\$2.24 (-\$6.30, \$1.81) p = 0.278	\$1.67 (-\$3.81, \$7.15) p = 0.549

	Bottom TS Tercile	**Middle TS Tercile**	**Upper TS Tercile**
Ads	\$19.81 (\$17.81, \$21.80)	\$34.04 (\$31.29, \$36.78)	\$40.27 (\$37.05, \$43.49)
No Ads	\$21.10 (\$18.78, \$23.42)	\$35.16 (\$32.14, \$38.18)	\$40.77 (\$37.09, \$44.45)
Δ Ads - No Ads	-\$1.29 (-\$4.40, \$1.81) p = 0.414	-\$1.12 (-\$5.23, \$2.98) p = 0.592	-\$0.50 (-\$5.57, \$4.56) p = 0.846

A.3 Additional Survey Details

In this section we first describe more details on the overall flow of the experiment on the user side and then provide more information on the survey response rates by group.

A.3.1 Additional Details on Survey Flow

For a user who was eligible to receive a survey and who logged in during our experimental window, the flow they would go through would be as follows: (i) they would be shown a prompt at the top of their newsfeed upon logging in asking them if they would like to participate in the survey; (ii) if they agreed, they would be asked a set of initial questions to determine if they would be willing to participate in a deactivation study; (iii) if yes, they would be asked our main valuation question for Facebook, and notified if they were randomly selected to receive a payment to deactivate; (iv) after that, they were asked some additional questions related to other research projects; and (v) finally, if they were randomly selected, they would be emailed their offer and be able to receive compensation for following through on the deactivation.

Below, we provide additional details on (ii), (iii), and (iv) that were not covered in the main text. We note that part of the data from this survey was also used in Brynjolfsson et al. (2023), and hence, the structure of the survey reflects multiple research projects.

Initial questions. If the user clicked 'Start Survey,' they would next be asked if they would be willing to participate in a deactivation study. Specifically, we asked: "Thinking about all the ways you use Facebook, would you be willing to stop using Facebook for one month if you were offered money in return?" Those who said "no" received a follow-up question asking why and then received a final opportunity to participate ("Would you like to learn more about the payment opportunity?"). Those who indicated they are willing to forego Facebook in exchange for money, or are at least willing to learn more in the follow-up question then proceeded to the Terms and Conditions.

The terms by which incentivized lotteries can be offered vary across countries. For example, some jurisdictions require different minimum ages or basic skills checks. After consenting and agreeing to the Terms and Conditions, the user was sent to the incentivized experiment.

Main valuation question. One implementation point on our valuation question concerns rounding. As mentioned in the main text, for countries that do not use USD, we converted the offer amounts to rounded values of the local currency. The amount of rounding depended on the level. Local currency amounts below 25 were rounded to the nearest integer; between 25 and 100 were rounded to the nearest 5; amounts between 100 and 500 were rounded to the nearest 25; between 500 and 1,000 to the nearest 50; between 1,000 and 5,000 to the nearest 100; between 5,000 and 10,000 to the nearest 500; between 10,000 and 100,000 to the nearest 5,000; and values about 100,000 to the nearest 20,000.

Additional questions. After soliciting users' answers to our incentivized experiment, we asked additional questions for other research projects and on basic demographics. These all occurred after we elicited users' answers to our Facebook valuation question, and users were not told about any subsequent questions, so we believe it is very unlikely these questions could have affected the earlier responses.

A.3.2 Survey Response Rates by Group

Table A.9 below contains the survey response funnel for the ads and no ads groups. We note the progression rates between the two groups are significantly different for progressing from being in the sampling frame to seeing a survey invitation and from seeing an invitation to starting the survey.

The largest percent-wise difference in progression rates for users in the ads versus no ads group is going from being in the sampling frame to seeing the survey invitation. The reason for this is (at least partly) due to Meta's survey infrastructure. When running a survey on platform at Meta, the researcher specifies ex ante a target number of survey starts. At the same time, the researcher selects the total number of users who are eligible to receive the survey and the time period over which the survey will field; the survey then runs until the target number of survey starts is met from the eligible population or the fielding period ends. In this case, we specified a target number of survey starts and an eligible set of users for each of the ads and no ads groups. Given the number of users who see ads is much larger than the number who do not, we were able to specify a relatively larger sample of eligible users for the ads condition to guard against a very low start rate that could hurt our power. (We also specified a larger number of survey starts for power as well.) This larger eligible population has the effect of mechanically increasing the denominator substantially for the progression rate in row two below for the ads group.

As mentioned in the main text, conditional on seeing the survey invitation, the response rates for the Facebook question for users in the ads and no ads groups were 3.7% and 4.7%, respectively. The largest (and significant) difference occurred between seeing the invitation and starting the survey. We suspect this difference is due to the fact that users in the no ads group tend to click slightly more on content. We note, however, that the magnitude is relatively small and both response rates are within the typical range seen in on-platform surveys (e.g., Alekseev et al., 2023).

Table A.9: Response funnel by group.

	Ads		No-Ads		
Funnel Step	**N**	**Progression Rate**	**N**	**Progression Rate**	***p*-value**
1) Sampling Frame	8,310,000	–	1,313,178	–	–
2) Saw Invitation	1,067,636	12.8%	287,118	21.9%	0.000
3) Started Survey	82,400	7.7%	28,076	9.8%	0.000
4) Answered FB Question	39,647	48.1%	13,436	47.9%	0.452

Note: The p-value in the last column is to test if the proportion of respondents is equal in each group (two-tailed).

A.4 Weighting Details

Our weighting strategy consists of three key building blocks: (1) design weights to account for differential probability of inclusion into the sample by country and experimental group, (2) user non-response weights to account for the probability of a user responding to the survey, (3) question non-response weights to account for the probability of a user who started the survey responding to the survey item in question. Below we go through how we calculated each of these for the respondents in our final dataset.

Design weights. For the design weights, the weight for user *i* from country *c* and experimental group *g* is given by

$$w_{cg}^{design} = \frac{\#\ monthly\ active\ Facebook\ users\ in\ country\ c\ and\ group\ g}{\#\ users\ from\ country\ c\ and\ group\ g\ that\ were\ eligible\ to\ receive\ the\ survey}$$

As the standard definition of design weights, they capture the number of units of our population of interest that each unit of our sample – those eligible to receive the survey – represents.

User non-response weights. Conditional on being eligible to receive a survey, only a subset of users actually started it. Our user non-response weights model the probability of starting the survey given a user was eligible to receive the survey as a function of observables. Specifically, we let the user non-response weight for user *i* in country *c* and experimental group *g* be:

$$w_{icg}^{user\ non-response} = 1 \,/\, \mathrm{P}(i \text{ started the survey} \mid i \text{ was eligible to receive the survey})$$

where P(*i* started the survey) is estimated separately for each county and experimental group and is a function of gender, primary phone operating system, whether the user has a profile picture, age (quartile bins), friend count (quartile bins), the number of days within the last 28 days that the user was active, an indicator for whether the user was active for all days within the last 28 days, and days since confirmed (i.e., tenure, in quartile bins). Following Sarig, Galili, and Eilat (2023), we estimate these probabilities using a regularized logistic regression with LASSO.

Question non-response weights. Finally, to correct for question non-response, we similarly inverse probability weight by the estimated probability a user who started the survey answered the Facebook valuation question. For user *i* in country *c* and experimental group *g* this is defined as:

$$w_{icg}^{question\ non-response} = 1/\ \mathrm{P}(i \text{ answered the Facebook valuation question} \mid i \text{ started the survey})$$

Similar to the user non-response weights, we estimate these weights separately for each country and experimental group and use the same logistic specification and set of right hand side variables. We include on the right hand side, however, an indicator for if they had answered they

would be willing to stop using Facebook for one month if they were offered money in return (an earlier question on the survey).

Finally, for estimates at the country-level, we multiply user and question non-response weights to obtain weights for each respondent, whereas for estimates where we pool data across countries, we multiply the three weights to obtain weights for each respondent.

References

Alekseev, G., Amer, S., Gopal, M., Kuchler, T., Schneider, J. W., Stroebel, J., & Wernerfelt, N. (2023). The effects of COVID-19 on US small businesses: evidence from owners, managers, and employees. *Management Science*, *69*(1), 7-24.

Sarig, T., Galili, T., & Eilat, R. (2023). balance--a Python package for balancing biased data samples. *arXiv preprint arXiv:2307.06024*.